Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 13, No. 2, 2025

Robotic Arm Visual Grasping Enhanced by an Improved U-Net Network

Zebin Liu [1], Batu Nasheng [2], Liangliang Lv [2], Haitao Wu [2], Wenbo Li [2], Yongbin Liu [2], Tianshi Qi [2]

[1] College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang, 310000, China
[2] Inner Mongolia Huomei-Hongjun Aluminum and Electricity Co., Ltd., Holingola, Inner Mongolia, 029200, China

Abstract: This paper proposes a robotic arm vision-based grasping method utilizing an enhanced U-Net architecture and an Efficient Channel Attention (ECA) mechanism, aimed at addressing the localization challenges of soft belts in industrial automation grasping. Traditional approaches suffer from grasping point localization errors when handling non-rigid objects due to deformation and boundary ambiguity issues. This study enhances U-Net's feature extraction capabilities by embedding an ECA module, combined with image moment theory and edge analysis, achieving precise extraction of target geometric features and optimized grasping point selection. Experimental results demonstrate that the improved U-Net_ECA model achieves mIoU scores of 93.59% and 88.47% on the training and validation sets, respectively, significantly outperforming conventional approaches. Furthermore, the proposed grasp point localization algorithm effectively resolves positioning errors caused by object deformation, validating its practicality and robustness in industrial settings.

Keywords: Robotic Arm Vision-Based Grasping; U-Net; Efficient Channel Attention (ECA); Flexible Objects; Grasping Point Optimization.

1. Introduction

With the continuous advancement of industrial automation, computer vision-based robotic arm grasping systems have become a vital component of modern intelligent manufacturing. This technology acquires environmental information through image sensors and employs intelligent algorithms to analyze spatial relationships between objects, providing precise operational guidance for robotic arms. This significantly enhances production efficiency and flexibility.

At the technical implementation level, vision systems typically employ high-precision industrial cameras to capture two-dimensional or three-dimensional data of target objects. Combined with advanced image processing algorithms, they extract key feature parameters, including core information such as the object's spatial pose and geometric dimensions[1]. This technology enables robots to accurately perceive their working environment, achieve autonomous decision-making, and perform precise operations. However, current technology still faces numerous challenges in practical applications. Further optimization is required in areas such as target recognition under complex conditions, adaptability to dynamic environments, and interference resistance.

Particularly in areas such as the robustness of image processing algorithms and the optimization of grasping points, significant technical bottlenecks remain to be overcome[2] Taking the task of gripping soft straps (woven bag handles) as an example, the automated handling of such flexible objects has long been a technical challenge. Traditional manual operations are not only inefficient but also suffer from issues like insufficient positioning accuracy. Addressing this situation, the development of an intelligent soft strap recognition and gripping system holds significant practical importance. In real-world production settings, soft straps often appear in randomly stacked configurations. Their deformable nature poses unique challenges for machine vision systems, necessitating advanced algorithms like deep learning and deformation modeling to achieve reliable grasping. Traditional rigid automation systems struggle to adapt to handling such unstructured objects. In contrast, vision-guided robotic arm systems, leveraging their superior environmental perception and adaptive capabilities, offer a novel technical pathway for intelligent processing of flexible objects.

2. Related Work

Research on visual robots can be traced back to the mid-to-late 20th century. Around the late 1960s, amid waves of technological innovation, the field of machine vision began to take shape. By the 1970s, breakthroughs in image processing technology improved target recognition accuracy and made the acquisition of depth information feasible for the first time[5].

Sekkat H et al.[6]proposed a neural inverse kinematics method based on deep deterministic policy gradients, combined with YOLOv5 object detection and inverse projection localization, to achieve autonomous grasping for multi-degree-of-freedom robotic arms. Compared to traditional methods, this approach demonstrates superior performance and accuracy. Li et al.[7]proposed a occlusion-aware grasping method (GOAL) based on binocular stereo vision. By inferring occlusion relationships, segmenting and localizing targets, and performing multi-pose estimation, it effectively addresses grasping challenges in multi-object scenarios. Its practicality was validated on an EPSON robotic arm. Hui et al.[8]proposed a grasp detection algorithm integrating 2D images with 3D point clouds. By generating 2D prediction boxes via the SSD network and optimizing boundary accuracy, combined with an improved PointNetGPD selective sampling strategy, the approach significantly reduces computational time while enhancing grasp success rates, enabling efficient real-time multi-object grasping. Liu et al.[9]proposed the RGBGrasp method, which establishes geometric constraints through a pre-trained deep prediction model. By integrating hash encoding and proposal sampling strategies, it achieves 3D environmental perception

and precise grasping—including transparent/mirrored objects—with only a limited number of RGB views. This significantly enhances the algorithm's adaptability and execution efficiency in complex scenarios. Ilangovan P et al.[10]proposed an intelligent weed removal robot system based on CNN image classification. By identifying weeds through image processing and controlling a robotic arm for precise cutting, the system achieved an 80% recognition accuracy in testing, delivering a low-cost, efficient, and environmentally friendly automated weed removal solution.

3. Method

3.1. An Improved Model Based on U-Net

As a deep learning model based on an enhanced fully convolutional neural network (FCN), U-Net was first proposed by Ronneberger's team in 2015[11]. Originally designed for segmentation tasks in biomedical images, it has since become a widely adopted solution in the field of computer vision. This model adopts a symmetric encoder-decoder architecture: the encoder module progressively extracts high-level semantic features and reduces spatial resolution through convolution and pooling operations; the decoder module restores resolution by utilizing deconvolution and cross-layer feature fusion. Its name originates from the similarity between the network topology and the shape of the letter "U". U-Net's core innovation lies in introducing cross-level skip connections—channelwise concatenating multi-scale features from each encoder stage with corresponding decoder layers. This design enhances the model's understanding of global contextual information while significantly improving the restoration of local details, ultimately achieving systematic improvements in segmentation accuracy.

Although deep learning models represented by U-Net have achieved remarkable results in image segmentation, they still exhibit several limitations. First, such models primarily rely on color and texture features for segmentation, paying insufficient attention to the geometric shape characteristics of target objects. This results in low edge localization accuracy and inadequate utilization of boundary information. Second, constrained by their shallow network depth, U-Net models suffer from insufficient feature extraction capabilities, making it difficult to effectively address the issue of blurred segmentation boundaries[12]. To address these shortcomings, this study proposes embedding an Efficient Channel Attention (ECA) mechanism within the U-Net architecture. This mechanism employs a lightweight local cross-channel interaction strategy to adaptively recalibrate channel feature responses without significantly increasing computational complexity, thereby effectively enhancing the model's ability to extract key features[13]. Compared to traditional attention mechanisms, the ECA module offers advantages in fewer parameters and higher computational efficiency, better balancing model performance and computational cost. This enhances segmentation accuracy and robustness in complex scenarios. The specific network architecture is illustrated in Figure 1.

This network adopts an end-to-end symmetric architecture. First, the encoder module employs a "convolution-attention-pooling" cascaded architecture. Each downsampling unit incorporates two consecutive 3×3 convolutional layers for feature extraction, followed by an embedded efficient channel attention (ECA) module. This module utilizes 1×1 convolutions to facilitate cross-channel information exchange, enabling adaptive learning of channel weights. Downsampling is then completed via 2×2 max pooling, with dynamically adjusted dropout rates based on network depth to suppress overfitting. Second, the decoder innovatively employs an "upsampling-feature concatenation-dimension reduction" workflow. After restoring feature map dimensions via 2×2 upsampling, it concatenates these with corresponding encoder-level feature maps (block copying) through skip connections. Channel redundancy is then reduced using 1×1 convolutions. Finally, the entire network achieves precise matching of feature scale and channel count through parametric design. Encoder feature map dimensions decrease by 2^n-fold while channel counts increase by 2^n-fold. The decoder performs feature reconstruction via a symmetric structure, ultimately outputting a single-channel segmentation mask through a 1×1 convolution.

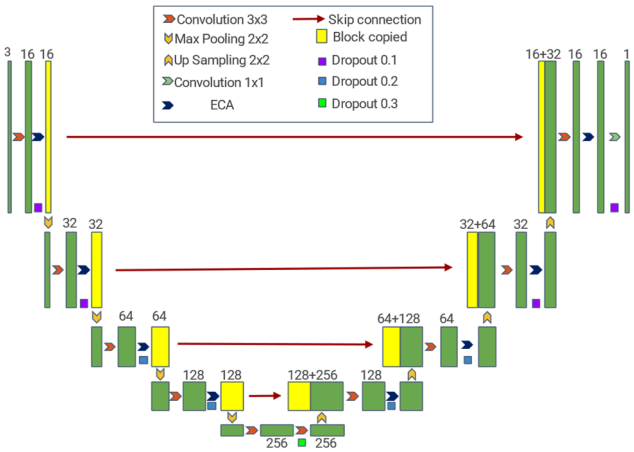

Figure 1. Improved U-Net Network Architecture

The internal structure of the ECA module is shown in Figure 2. First, the input feature map undergoes a Global Average Pooling (GAP) operation, compressing the spatial information of each channel into a 1×1×C channel statistical vector, thereby decoupling the channel dimension from the spatial dimension. Subsequently, to prevent information loss caused by dimensionality reduction in traditional channel attention mechanisms, this vector undergoes local cross-channel interaction through a 1×1 convolutional layer. The kernel size is adaptively determined based on the number of channels to balance feature interaction capabilities across different channel scales. Next, the convolutional output is normalized via a Sigmoid activation function, generating attention weights for C channels. Finally, the weights are fused with the original input feature map through per-channel multiplication, dynamically enhancing key channel features to produce the weighted feature map. ECA employs global average pooling and 1×1 convolutions with adaptive kernel sizes to adaptively learn channel importance without excessive computational overhead. It particularly amplifies subtle feature channel responses in soft-edge regions (e.g., object edges, blurred boundaries, or grayscale gradient areas) while suppressing interference from background noise channels. Combined with U-Net's fusion of shallow-layer details and deep-layer semantics through skip connections, ECA guides the model to more accurately capture feature differences in soft-edge regions. This mitigates the edge blurring and category confusion issues traditionally encountered in soft-edge segmentation by U-Net, ultimately enhancing segmentation completeness and boundary

localization accuracy in soft-edge areas.

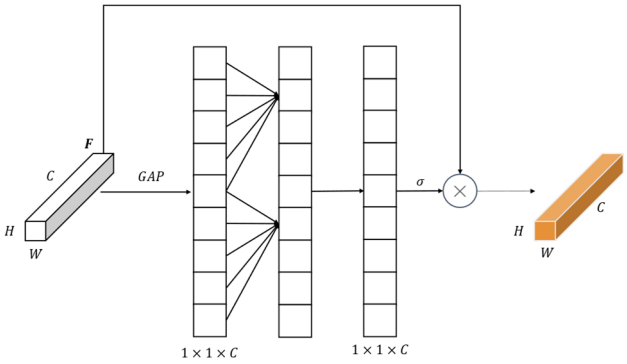

Figure 2. ECA module structure

3.2. Object Pose Estimation Methods

Traditional methods for calculating the center of gravity based on image moments exhibit significant limitations in locating the center of gravity of flexible objects such as soft belts. Due to their highly deformable nature, soft belts undergo nonlinear deformation under external forces during grasping, causing the center of gravity calculated from binary image pixel distributions to deviate significantly from the actual physical center of gravity. To this end, this paper proposes an object pose estimation method based on image moments and edge analysis, achieving systematic extraction of geometric features in the target region through a multi-stage processing workflow. The method first locates the centroid position of connected regions, then identifies boundary feature points, and finally obtains the dominant direction angle of the region using line detection techniques.

A centroid localization method based on image moment theory abstracts the pixel gray values of discrete binary images into a mass density distribution function, achieving a mathematical representation of target geometric features through spatial moment operations[14]. For a given binary image function $I(x,y) \in \{0,1\}$,its $(p+q)$ th-order geometric moment can be expressed in discrete form as:

$$m_{pq} = \sum_{x=1}^{W} \sum_{y=1}^{H} x^p y^p I(x,y) \tag{1}$$

In the equation, the zero-order moment m_{00} represents the total number of pixels in the target region, while the center-of-gravity coordinates are determined by the ratio of the first-order moment to the zero-order moment:

$$c_x = \frac{m_{10}}{m_{00}}, c_y = \frac{m_{01}}{m_{00}} \tag{2}$$

The computational model determines the region to be invalid at time $m_{00} = 0$, at which point the system concludes that no valid center of mass exists within the target area.

Object boundary extraction employs a topology-preserving contour tracking algorithm. This method precisely reconstructs the discrete point set of the target object's closed contour by establishing a rigorous neighborhood relationship model[15]. Given a set of contour points $C = \{(x_i, y_i)\}_{i=1}^{N}$ and their centroid coordinates (c_x, c_y), formulate the following minimization problem:

$$(c^*, y^*) = \arg\min \sqrt{(x_i - c_x)^2 + (y_i - c_y)^2} \tag{3}$$

Here, (c^*, y^*) denotes the boundary point on the contour closest to the center of mass. This algorithm ensures the topological correctness of boundary extraction through octant connectivity analysis, with a time complexity of O(N). It achieves sub-pixel-level boundary localization accuracy while maintaining computational efficiency.

The primary direction angle estimation employs a hybrid strategy combining Canny edge detection with the Hough transform. This method first identifies regions of interest (ROIs) within the target area's neighborhood, then applies the Canny edge detection operator to compute gradients and extract edges. A dual-threshold suppression strategy is used to filter out high-confidence edge contours. Subsequently, the preprocessed ROI undergoes Probabilistic Hough Transform (PHT) for line segment detection. By configuring parameters such as minimum segment length, maximum gap, and accumulator threshold, significant linear features are effectively extracted. Finally, all detected lines undergo quantitative length analysis. The longest line is selected as the primary direction reference for the target region. Based on its endpoint coordinates, the angle between this line and the image's horizontal axis is calculated, enabling precise estimation of the primary direction angle.

4. Experiment

4.1. Dataset and Experimental Settings

The experimental data used in this study were collected from an aluminum production site in Hohhot, Inner Mongolia. A total of 187 industrial-grade grayscale images were obtained, primarily capturing packaging bags for alumina powder. A typical sample is shown in Figure 3.

The image acquisition system consists of a Hikvision MV-CH100-60GM 10-megapixel industrial camera paired with an MVL-KF0618M-12MPE professional lens. The system supports a resolution of 4096×2460 pixels and 8-bit grayscale depth. During capture, the camera is vertically mounted 1.5 meters above the flexible bulk bag to ensure consistent imaging perspective. The dataset randomly divides all 187 images into training and test sets at a 9:1 ratio.

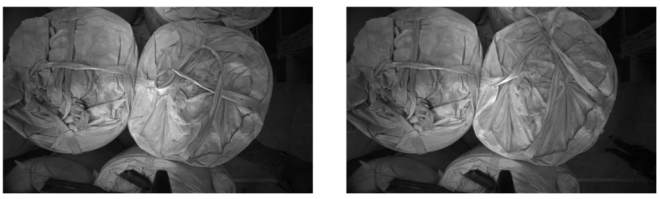

Figure 3. Partial Data Display

4.2. Analysis of Experiment Results

4.2.1. Algorithm Comparison Experiment

This study employs an improved U-Net architecture with VGG16 as its backbone. Through a two-stage training strategy (50 rounds of frozen training followed by 200 rounds of full-parameter optimization), the model was trained using the Adam optimizer (initial learning rate 1e-4, cosine annealed to 1e-6) with a 512×512 input size and a batch size of 4. The entire process was GPU-accelerated and utilized randomly initialized parameters.

As shown in Table 1, comparative experiments between the U-Net model and DeeplabV3+ model demonstrate that the U-Net model significantly outperforms DeeplabV3+ on both the training and validation sets. Its mIoU, mPA, and mPrecision on the training set reached 88.70%, 93.31%, and 94.05%, respectively, while on the validation set, the corresponding metrics were 86.17%, 90.96%, and 93.15%. In contrast, the

DeeplabV3+ model achieved mIoU, mPA, and mPrecision of 74.21%, 79.44%, and 87.47% on the training set, and 71.62%, 75.86%, and 87.17% on the validation set. U-Net demonstrates superior generalization performance and segmentation accuracy in soft-band semantic segmentation tasks.

Table 1. Results of Soft Band Segmentation Using Different Segmentation Models

Model	Training (%)			Validation (%)		
	mIoU	mPA	mPrecision	mIoU	mPA	mPrecision
DeeplabV3+	74.21	79.44	87.47	71.62	75.86	87.17
U-Net	88.70	93.31	94.05	86.17	90.96	93.15

4.2.2. Melting Experiment

The ablation results are shown in Table 2. The U-Net_ECA model, which incorporates the ECA(Efficient Channel Attention) module, outperforms the baseline U-Net model on both the training and validation sets. Specifically, on the training set, U-Net_ECA achieved mIoU, mPA, and mPrecision of 93.59%, 96.41%, and 96.77%, respectively, representing improvements of 4.89%, 3.10%, and 2.72% over U-Net. On the validation set, its mIoU, mPA, and mPrecision were 88.47%, 91.89%, and 95.28%, respectively, representing improvements of 2.30%, 0.93%, and 2.13% over U-Net. This demonstrates that the ECA module effectively enhances the model's channel feature extraction capability.

Table 2. u-net ablation experiment comparison

Model	Training (%)			Validation (%)		
	mIoU	mPA	mPrecision	mIoU	mPA	mPrecision
U-Net	88.70	93.31	94.05	86.17	90.96	93.15
U-Net_ECA	93.59	96.41	96.77	88.47	91.89	95.28

4.2.3. Grab Point Comparison

As shown in Figure 4, in the soft belt grasping scenario, the proposed algorithm effectively addresses the issue of grasping point positioning deviation caused by the irregular shape of the soft belt. The original image on the right shows that when the centroid of the soft belt region (blue mark) is directly used as the grasping point (red dot mark), the non-rigid characteristics of the soft belt's shape may cause the centroid position to deviate from the actual physical contour. In extreme cases, the grasping point may even fall outside the soft belt (indicated by the blue circumscribed circle). In contrast, the optimized result image on the left demonstrates successful correction and precise localization of the grasping point onto the actual physical contour of the soft belt through integrated analysis of its geometric features and physical properties. Furthermore, the nearest neighbor search results marked by red lines validate the spatial correlation between the grasping point and the soft belt's surface.

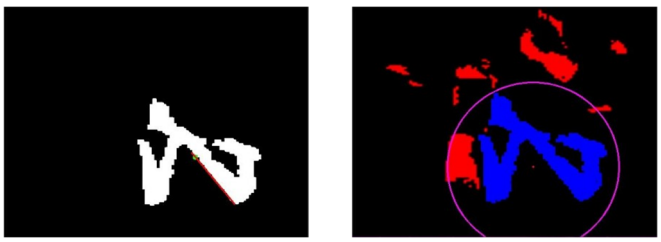

Figure 4. Comparison of Grab Point Locations

5. Experiment

This paper addresses the positioning challenge of soft belts in robotic arm vision-based grasping by proposing a solution based on an improved U-Net and an Efficient Channel Attention (ECA) mechanism. By embedding the ECA module, the feature extraction capability of the U-Net model is significantly enhanced. Combined with image moment theory and edge analysis methods, it achieves precise extraction of target geometric features and optimization of grasping points, demonstrating its practicality and robustness in industrial scenarios.

References

[1] Chiu YJ, Yuan YY, Jian SR. Design of and research on the robot arm recovery grasping system based on machine vision. Journal of King Saud University-Computer and Information Sciences. 2024 Apr 1;36(4):102014.

[2] Yin X, Chen Y, Guo W, Yang Z, Chen H, Liao A, Yao D. Flexible grasping of robot arm based on improved Informed-RRT star. Chinese journal of engineering. 2025 Jan 25;47(1):113-20.

[3] Afzal W, Iqbal S, Tahira Z, Qureshi ME. Gesture control robotic arm using flex sensor. Applied and Computational Mathematics. 2017 Jan;6(4):171-6.

[4] Zhou J, Chen S, Wang Z. A soft-robotic gripper with enhanced object adaptation and grasping reliability. IEEE Robotics and automation letters. 2017 Jun 16;2(4):2287-93.

[5] Domae Y. Recent trends in the research of industrial robots and future outlook. Journal of Robotics and Mechatronics. 2019 Feb 20;31(1):57-62.

[6] Sekkat H, Tigani S, Saadane R, Chehri A. Vision-based robotic arm control algorithm using deep reinforcement learning for autonomous objects grasping. Applied Sciences. 2021 Aug 27;11(17):7917.

[7] Li L, Cherouat A, Snoussi H, Wang T. Grasping with occlusion-aware ally method in complex scenes. IEEE Transactions on Automation Science and Engineering. 2024 Aug 1.

[8] Hui NM, Wu XH, Han XW, Wu BJ. A robotic arm visual grasp detection algorithm combining 2D images and 3D point clouds. Applied Mechanics and Materials. 2024 Mar 5;919:209-23.

[9] Liu C, Shi K, Zhou K, Wang H, Zhang J, Dong H. Rgbgrasp: Image-based object grasping by capturing multiple views during robot arm movement with neural radiance fields. IEEE Robotics and Automation Letters. 2024 May 2;9(6):6012-9.

[10] Ilangovan P, Ilangovan, K. Meena, M. S. Begum: Weed Easy Extraction Using Deep Learning With a Robotic System,

Proceedings of the International Conference on Physics and Engineering (Virtual Conference, December 1, 2024), Vol. 2923, p. 012003.

[11] O. Ronneberger, P. Fischer, T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (Munich, Germany, October 5-9, 2015). Vol. 9351, p. 234.

[12] Srinivasan S, Durairaju K, Deeba K, Mathivanan SK, Karthikeyan P, Shah MA. Multimodal biomedical image segmentation using multi-dimensional U-convolutional neural network. BMC Medical Imaging. 2024 Feb 8;24(1):38.

[13] Pacal I, Celik O, Bayram B, Cunha A. Enhancing EfficientNetv2 with global and efficient channel attention mechanisms for accurate MRI-Based brain tumor classification. Cluster Computing. 2024 Nov;27(8):11187-212.

[14] Teague MR. Image analysis via the general theory of moments. Journal of the optical society of America. 1980 Aug 1;70(8):920-30.

[15] Liu G, Li H, Yang L. A topology preserving method of evolving contours based on sparsity constraint for object segmentation. IEEE Access. 2017 Sep 18;5:19971-82.